CMPT 741/459 Course Project

100 points, due at 11:59 pm, December 4, 2016 (Friday)

Please use font size no smaller than 10 points. Your project report and results should be submitted online through the assignment submission site in an archive file. No late submission will be taken or graded.

Please put your student-id number and whether you are a CMPT 459/741 student at the beginning of your project report. For CMPT 741 students, please identify yourself whether you are a thesis-based student.

Data Set

The data set used in this project is derived from the first four mini assignments in this course. We encode a word appearing in the essays using a unique positive integer. The essays are preprocessed properly, that is, they are stemmed and lemmatized, and stop-words are removed[1]. Every student is also encoded by a unique positive integer.

The data set consists of three files.

- `Collaboration.txt`. This file records the collaboration relations among all students. Each line contains 3 integers in the following format

  $$u, v, d$$

  which means student u and student v collaborated in a mini assignment and produced document d. It is possible that $u = v$. In such a case, the document d was submitted by only one student.

- `DocumentWords.txt`. This file contains the essays. There are in total N lines in this file, where N is the number of essays. The i-th line lists the words used in the i-th document. Specifically, the format of a line is

  $$d_i, k_i, w_{i_1}, w_{i_2}, \ldots, w_{i_{k_i}}$$

  where d_i is the id of the i-th essay (corresponding to d in the file `Collaboration.txt`), k_i is the number of different unique words in this essay, and $w_{i_1}, w_{i_2}, \ldots, w_{i_{k_i}}$ are the words.

- `Vocabulary.txt`. This file lists the words used in the file `DocumentWords.txt`.

The data set can be downloaded at `http://www.cs.sfu.ca/CourseCentral/741/jpei/slides/ProjectData.zip`.

[1] If you do not know what those are, it does not matter. Please just assume that the data is properly cleaned.

Task for CMPT 459 Students

Your task in this project is to partition all essays into 4 groups exclusively. That is, every essay belongs to one and only one group. The essays belonging to the same mini assignment should be clustered together.

Submission Requirements

- Please submit a text file named by your student id (e.g. `301226685.txt`) that contains N lines (N is the number of essays). The format of the i-th line should be

  $$d_i, c_i$$

  where d_i is the essay id, and c_i is the cluster that d_i belongs to. Please note that $c_i \in \{1, 2, 3, 4\}$.

  You are NOT expected to have the cluster ids correspond to the mini assignment ids. That is, it is completely fine if in your answer cluster 2 corresponds to mini assignment 4.
- Please also submit a project report of NO MORE THAN 2 PAGES describing how you tackle the task in this project. The report should at least include the following parts: methods you adopt, experiments you conduct and the results, analysis of the experimental results and references.

Evaluation

Your project will be evaluated in the following 2 parts.

Clustering Results (50%)

We use the ground truth to evaluate your results.

For a pair of essays (d_i, d_j) $(i \neq j)$, your submission is right if (1) d_i and d_j belong to the same mini assignment and in your result $c_i = c_j$; or (2) d_i and d_j belong to different mini assignments and in your clustering $c_i \neq c_j$. Let $I(i, j)$ be the correctness of your result on the pair of essays (d_i, d_j), which takes 1 if your result is correct on (d_i, d_j) and 0 otherwise. Then, your result will be evaluated by the accuracy defined as

$$Accuracy = \frac{2}{N(N-1)} \sum_{(i,j)} I(i, j) \tag{1}$$

Project Report (50%)

Your report will be graded according to its clarity, appropriateness, soundness and originality.

Early Validation

On November 28, 2015 (Saturday), we will provide you an executable program that can give you the accuracy of your results based on a random sample of 20% of the essay pairs. However, the sample itself will not be provided. Please keep in mind that the accuracy estimated by this program does not necessarily equal your final accuracy based on the complete ground truth.

Task for CMPT 741 Students

The list of words in each essay provided in the file `DocumentWords.txt` is in fact incomplete. For each essay, we randomly removed some words from the list. Your task is to predict those missing words.

Submission Requirements

- For each document, please provide a list of 5 words that you think are most likely removed by us. Please submit a text file named by your student id (e.g. `301226685.txt`) where each line is in the following format

  $$d_i, w_1^i, w_2^i, w_3^i, w_4^i, w_5^i$$

  where d_i is the essay id, $w_1^i, w_2^i, w_3^i, w_4^i, w_5^i$ are the 5 words you recommend.

  Please note that the order of the 5 words really matters in evaluating your prediction. Please sort the 5 words in the descending order of your confidence on their belonging to essay d_i.

- Please also submit a project report of NO MORE THAN 2 PAGES describing how you tackle the task in this project. The report should at least include the following parts: methods you adopt, experiments you conduct and the results, analysis of the experimental results and references.

Additional Requirement for Thesis-based Student Only

Please investigate the usefulness of the file `Collaboration.txt` in this project. Does it help in the prediction? If it does, how do you incorporate the information contained by it in your prediction model? If not, why? Please discuss these issues and provide corresponding experiment results in your project report. You can have an extra page on this part.

Evaluation

Your project will be evaluated in the following 2 parts.

Prediction Results (50%)

The precision of the first k words you recommend for essay d_i is defined as

$$P_k^i = \frac{\sum_{j=1}^{k} I(w_j^i \in d_i)}{k}$$

where $I(x)$ is an indicator function, which takes 1 if x is true and 0 otherwise. Accordingly, we can define the measure $MAP^i@5$ as

$$MAP^i@5 = \frac{\sum_{k=1}^{5} P_k}{5}$$

Your prediction of the missing words is evaluated by the average $MAP@5$ value of all documents, that is,

$$ave(MAP@5) = \frac{\sum_{i=1}^{N} MAP^i@5}{N}$$

Project Report (50%)

Your report will be graded according to its clarity, appropriateness, soundness and originality.

Early Validation

On November 28, 2015 (Saturday), we will provide an executable program that can give you the average $MAP@5$ value of your prediction based on a random sample of 20% essays. However, the sample itself will not be provided. Please keep in mind that the estimate by this program does not necessarily equal your final $MAP@5$ based on all essays.